From GIS to Graphical Representation for Maintaining Connectivity of Wastewater Network Elements

Omar Et-Targuy, Carole Delenne, Salem Benferhat, Ahlame Begdouri, Thanh-Nghi Do, Truong-Thanh Ma

From GIS to graphical representation for maintaining connectivity of wastewater network elements

Omar Et-Targuy[1,2,3,5,*], Carole Delenne[3,5], Salem Benferhat[1], Ahlame Begdouri[2], Thanh-Nghi Do[4], and Truong-Thanh MA[4]

[1] CRIL, CNRS UMR 8188, University of Artois, France.
[2] LSIA, Université Sidi Mohamed Ben Abdellah (USMBA), Fez, Morocco.
[3] HSM, University of Montpellier, CNRS, IRD, Montpellier, France
[4] CICT, Can Tho University, Vietnam
[5] Inria, Team Lemon, Montpellier, France

Abstract. Wastewater network management is a critical aspect of ensuring public health and environmental sustainability. Geographic Information Systems (GIS) are widely employed for managing and analyzing wastewater network data. A Shapefile is a model for representing vector data in GIS. In this model, different databases are used to store various types of geometric data, including points, lines, and polygons. For instance, the components of wastewater networks are stored in different databases. However, this approach presents difficulties in accurately depicting the connectivity among the wastewater network components that are physically interconnected in reality. To address this issue, this paper discusses the limitations of the separate databases approach for representing wastewater networks in Shapefile and proposes a novel graph-based approach. In this approach, each component of the network, such as manholes, structures, pumps, etc., is represented as a node in the novel graph, while the pipes represent the connections between them. By adopting the proposed representation, the real interconnected nature of the wastewater network can be effectively captured and visualized. The validation of this approach, using five real datasets, confirms its ability to connect the various components of the wastewater network via a graph-based representation.

Keywords: Wastewater networks · Geographical Information System· Shapefile · Graph-based representation · Connectivity.

1 Introduction

Efficient wastewater management plays a vital role in ensuring public health and environmental sustainability [11]. Geographic Information Systems (GIS) have become a popular tool for managing and analyzing geographic data associated with wastewater networks [16, 15]. Different models are used to store vector data in GIS. Shapefile is a non-topological model that allows storing different geometries in different databases without storing the topological relation

between them. However, the approach of storing each component of wastewater networks in separate databases using Shapefiles lead to imprecision, especially when it comes to representing the connections between these components. In reality, wastewater networks are interconnected systems where various components work together to transport wastewater from its source to treatment plants. For example, manholes serve as access points to the pipes, facilitating maintenance and inspection activities. These connections between different components are crucial for the proper functioning of the network.

While in Shapefiles, storing different components in different databases allows the efficient management of individual components, it lacks the ability to capture and represent the connections between them due to the non topological characteristic of Shapefile model. Indeed, the manual insertion of wastewater components data in a GIS may lead to slight shifts in position, resulting disconnected objects. Furthermore, the use of different types of geometries (points, lines and polygons) to represent real-world objects may also introduce imprecision in terms of position and connectivity, since selecting the most appropriate geometry to represent real-world objects is not a straightforward process. As a result, over time, as updates and modifications are made to the network, the databases become unsynchronized, leading to a loss of the real connectivity and relationships between the components. This is a challenging problem since the correspondence between nodes and pipes is not perfect due to the imprecise positioning of components, while some components are isolated from the network, further complicating the connectivity problem.

To address the challenges related to the representation of wastewater networks, a comprehensive and integrated approach is needed. A graph-based representation is very natural for capturing the interconnected nature of the network, the connectivity between nodes and edges being a fundamental characteristic of graphs. In this approach, each component, including manholes, pumps and structures, is treated as a node, while the connections between them are represented as edges, symbolizing the pipes in wastewater networks. In this approach, dummy nodes are introduced to ensure connectivity between the various components and to represent the missing nodes in wastewater networks. However, capturing the connections between the various components is a major problem, due to the imprecision associated with the Shapefiles representation approach and its non topological characteristic.

This paper is an extended version of work published in [6]. We extended our previous work by enhancing the state of the art, improving the graph construction approach and extending the experiments. It is organized as follows. The overview of GIS, vector data models, and wastewater network data is provided in Section 2, where we discuss the limitations of representing them in a Shapefile using the traditional separate databases approach, highlighting the resulting connectivity issues. Section 3 presents the graph-based algorithm, which transforms the shapefile representation into a graph-based representation. In Section 4, experiments and results from applying the graph-based algorithm to real-world wastewater network data are presented, with a focus on highlighting the

algorithm's advantages and limitations. Finally, Section 5 concludes the paper by summarizing the benefits of the graph-based representation and its implications for wastewater network management.

2 Wastewater networks in GIS

2.1 Geographical Information Systems (GIS)

According to [3], GIS is a powerful set of tools for collecting, storing, retrieving, transforming, and displaying spatial data from the real world. GIS manages both attribute and geospatial data. Attribute data is organized in tabular form, while geospatial data is stored in raster or vector formats. Raster data comprises a pixel matrix, where each pixel holds one or several values, stored in different bands, such as RGB (Red, Green, Blue) or environmental data like temperature or elevation. Vector data, defined by XY coordinates, is categorized into points, lines and polygons. Points represent small or abstract features, lines denote linear elements with measurable lengths such as road networks, and polygons depict boundaries and areas of features such as buildings, cities boundaries or lakes.

Vector data can be structured in many different ways [4]. The three most common data models are :

1. The topological data model: This model is characterized by the inclusion of topological information within the dataset, as the name implies. Topology is a set of rules that model the relationships between neighboring points, lines, and polygons and determines how they share geometry.
2. The spaghetti data model [5]: In the this model, each point, line, and/or polygon feature is represented as a string of X, Y coordinate pairs (or as a single X, Y coordinate pair in the case of a vector image with a single point) with no inherent structure. The topological information are not explicitly encoded in this model.
3. The Esri Shapefile model [21]: This is a non-topological model, meaning it does not store topological information between different geometries. It is one of the most common GIS vector data formats and is compatible with the majority of software platforms. A Shapefile is a straightforward file format that stores the geometry and attribute information of the geometries in a tabular format. Hence, a single Shapefile will be referred to as a database in this context.

In the research context, GIS has been used in several works. In [17], a novel route planning methodology addresses traditional impedance modeling challenges by introducing realistic variables into GIS. The proposed impedance model, based on the hierarchical analytical process, successfully enhances route planning accuracy. The study suggests exploring complementary techniques like the network analytical process for further GIS methodology improvement. In the study [2], integrated remote sensing and GIS were used to identify areas of earthquake potential. The study introduced a digital ranking system for the

integration of GIS data, resulting in a more accurate and detailed Earthquake Potential Index (EPI) map instead of traditional earthquake risk maps. In [9], a study of integrated GIS in developing high-resolution global climate surfaces using monthly data from 1950-2000 is presented. Employing advanced interpolation techniques, the research quantified uncertainties and highlighted the value of these surfaces in providing insights into regional climate variations despite low station density.

2.2 Wastewater network

Wastewater networks comprise a variety of interconnected components that collectively ensure the efficient and safe transportation of wastewater from its sources to treatment plants [19]. Some of the essential components of wastewater networks include:

- Pipes: sewer pipes are the primary components of a wastewater network. They are used to transport wastewater from its source, such as residential, commercial or industrial buildings, to the treatment plants.
- Manholes: manholes are access points in the wastewater network that allow for maintenance and inspection activities. They are usually constructed at regular intervals along the sewer pipes. Manholes provide entry points for workers to access the sewer pipes and perform tasks like cleaning, repairs, and inspections.
- Pumps: pumps are used to increase the pressure and flow of wastewater, especially in areas where gravity flow is not possible.
- Devices: Various devices play crucial roles in the proper functioning of wastewater networks. These include storage basins, plugs or stoppers, chambers, flush tanks, and siphons.

One characteristic of wastewater network data is the geographical information associated with each component, representing its position on the Earth. To store, visualize and manage this spatial data, a GIS is commonly employed. A GIS allows for the integration, analysis and interpretation of spatial data within the context of wastewater network management.

Wastewater network data is presented in vector format. Consequently, two main geometries are employed to depict the network components in GIS, as shown in Figure 1, representing a section of the wastewater networks in the city of Montpellier, France. The first geometry is the point geometry, representing specific locations such as manholes, structures, pumps and other components. The second geometry is the line geometry, which illustrates the pipes constituting the network infrastructure. The line geometry is used to represent pipes only, all other components are represented by point geometry. The vector data model used for representing wastewater network data is the Shapefile model, which is a non-topological model. This means that the topological relationships between point geometry and line geometry are not explicitly represented. Throughout this article, the representation of wastewater networks using the Shapefile model in GIS will be referred to as the GIS representation.

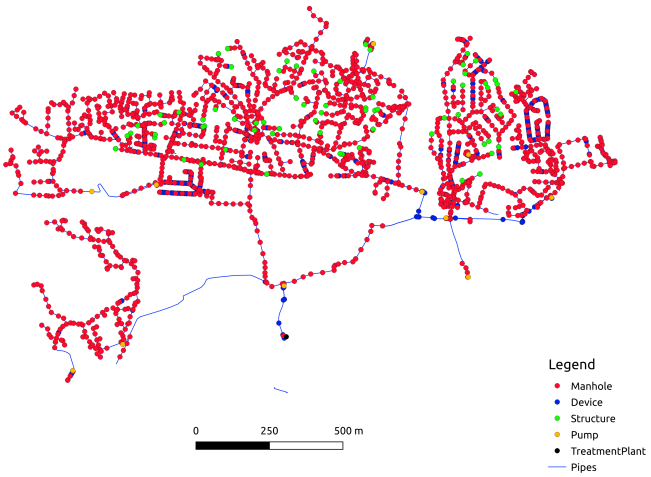


Fig. 1: Example of wastewater networks in GIS representation (Open data of Montpellier city [14]).

In addition to the geographical information, each component in the wastewater network is also characterized by attribute information. These attributes provide additional details and characteristics of the components, such as pipe diameter, material, flow capacity and other relevant data.

In a Shapefile model, the data related to each component of the wastewater network is typically stored in separate databases, as illustrated in Figure 1. Each component, such as pipes, devices, spillway, manholes, pumps, structures and treatment plants, has its own database, allowing for easy management and analysis of the data.

2.3 Graph-based representation

Graph-based representation is a versatile method adept at illustrating data interrelationships across diverse domains. In graph representation, nodes serve as placeholders for various entities or abstract ideas, while edges between them represent various relationships such as similarity or connectivity, depending on the context. For instance, in the context of previous studies, a pioneering graph-based spatial approach is introduced in [7], employing minimal planar graphs from Delaunay triangulations to effectively assess habitat connectivity. This innovation facilitates wildlife habitat network analysis and visualization, aiding in conservation decisions. In [18], the authors conduct a systematic survey of graph-based text representation, highlighting its superiority over the bag-of-words model in capturing structural and semantic information. The use of graph-based representation leads to enhanced performance in information retrieval, term weighting and ranking across various text applications. In [12], the

authors address problems posed by multi-view video sharing and propose a more efficient way of representing and compressing 3D information. Using a method called graph-based representation, the focus is on connecting important pixels between views for efficient compression. Additionally, in [13], a novel graph-based approach for multi-view image representation is proposed. This method utilizes graph connections to describe pixel proximity in 3D space, achieving a compact and controllable representation well-suited for coding and reconstructing multiple views. In [10], a knowledge graph-based data representation is proposed for Industrial Internet of Things-enabled manufacturing. In [22] a global graph for modeling social interactions among pedestrians has been investigated. These research studies demonstrate a strong interest in graph-based representation, which they exploit to capture connectivity or similarity between abstract entities or ideas. This highlights the versatility and effectiveness of graph-based methods in different fields.

2.4 Problem statement

In the real world, all of wastewater network's components are interconnected and have spatial relationships with each other; e.g., manholes serve as access points to pipes, and pumps are connected to specific pipes within the network. Therefore, it is crucial to establish and maintain the appropriate connections between these components in terms of representation.

As the amount of wastewater network data grows, it becomes increasingly important to update the nodes and pipes that represent these networks in databases. However, ensuring that these databases are properly synchronized with each other is challenging. Imprecision in the representation of the wastewater network arises when the connections between the components are not accurately captured or updated. For example, Figure 2 illustrates some of the issues that can arise with this type of representation. In the figure, some nodes are not connected to any pipes, which is impossible in reality. Additionally, in some cases, some pipes don't have a node at one of its extremities or both. Many other potential problems arise when the nodes and pipes of a wastewater network are not properly synchronized.

To address these issues, a graph-based representation is one of the most appropriate methods for ensuring connectivity between different components of the network. By using a graph structure, nodes representing manholes, structures, pumps and other components can be connected by edges representing pipes. This graph-based approach accurately reflects the interconnected nature of the wastewater network. Additionally, the graph-based representation helps to complete data by addressing missing nodes in the GIS representation, ensuring a more comprehensive view of the wastewater network. Moreover, by retaining all the original information, it enables a return-back to the GIS representations whenever needed. It also facilitates the integration of new information into the graph, making it a flexible and scalable solution for wastewater network management.

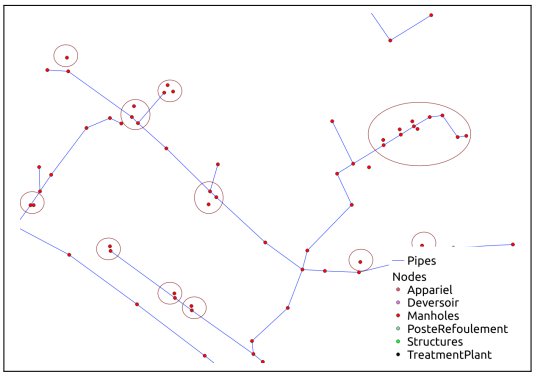


Fig. 2: Example of connectivity issues: mainly nodes not connected to a pipe and missing nodes at pipes' extremities.

In the following section, the details of the graph-based representation and its benefits in maintaining the connectivity and integrity of wastewater network data are presented.

3 Graph-based representation of wastewater networks

3.1 Domain constraints

A diverse array of stakeholders play integral roles in the establishment and operation of wastewater networks. Recognizing the need for seamless collaboration and efficient data sharing among these stakeholders, it becomes imperative to establish a comprehensive set of technical and topological constraints. Adherence to these domain constraints is crucial during the graph construction process to guarantee the reliability and efficiency of wastewater network management. Based on our study of the technical documents related to wastewater networks [1] of Montpellier Mediterranean Metropolis zone in France , some of the essential domain constraints include:

- Duplicated Objects: It is impossible to encounter two components of the same type at the same location in reality. For example, the presence of two manholes at identical positions is impossible.
- Data Model Specification: Wastewater pipes are visually represented by lines or poly-lines (arcs), while equipment points are symbolized by nodes.
- Uninterrupted Network Design: Arcs must seamlessly connect end-to-end, each commencing and concluding with a node.
- Coordinates Consistency Requirement: The coordinates of a node must precisely match those at the end of the corresponding pipe.
- Strategic Placement of Pipe Junctions: Nodes are strategically introduced at intersections, even if they do not belong to a specific component.

- Flow Direction Alignment: The network is systematically drawn in the direction of effluent flow, proceeding from upstream to downstream.
- Nominal Diameter Standard: Pipes are mandated to possess a nominal diameter equal to or exceeding 200 mm, encompassing private networks.

3.2 Formal representation

The inputs of the graph-based representation approach can be formally represented as follows:

- The Pipes Database PD is a collection of line geometries that represent sewer pipes PD $= \{p_1, ..., p_i, ..., p_k\}$. Each pipe p_i is identified by a unique identifier i, and the total number of pipes in the database is denoted by N_P.
- The Nodes Database ND is a collection of point geometries that includes all the point type components: manholes, structures, pumps, fittings, accessories, devices, spillway and treatment plants ND $= \{n_1, ..., n_j, ..., n_m\}$. Each element n_j has an attribute *type*, allowing us to identify the type of component. Each node n_j is identified by a unique identifier j, and the total number of nodes is denoted by N_N.

A simple graph G is a pair (V, E), where V is a finite set, and where E is a set consisting of 2-element subsets of V [8]. We will abbreviate the word "simple graph" as "graph" in this article. The algorithm generates a graph denoted as $G = (V, E)$ as its final output, where:

- V is a finite set of nodes that contains elements from ND and dummy nodes that will be added to ensure connectivity and complete missing nodes.
- $E \subseteq V \times V$ is a finite set of edges that represents connections between two nodes, indicating that these connections correspond to the pipes from PD. Each edge in the graph represents the relationship between two nodes, reflecting the pipes that connect them in the wastewater network representation.

3.3 General approach

The aim of our approach is to construct a graph-based representation of wastewater networks in order to ensure connectivity between the various components.

Figure 3 illustrates the general approach to reach this goal. The inputs include a database of pipes and multiple databases of nodes. A pre-processing phase is implemented to remove all duplicated objects. Duplications may be encountered as a result of the manual insertion of data in GIS while creating or updating the network. Nodes and pipes are, in this case, characterized by identical positions and attributes; which is not allowed by the domain constraints (3.1). Furthermore, the presence of distinct databases for each component type considerably extends the time required to identify intersections between nodes and pipes. To streamline this process and enhance efficiency, we combine the node databases into a unified node database.

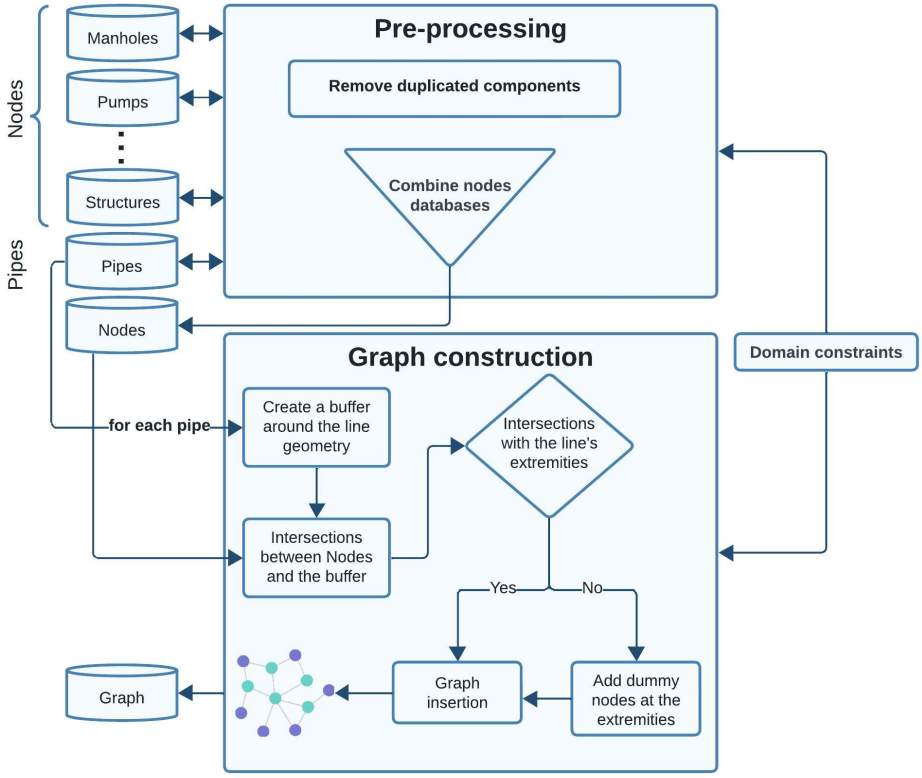


Fig. 3: General approach of graph construction from GIS pipes and nodes databases.

The graph construction algorithm starts by extracting the intersections between each pipe in the pipes database and the nodes in the nodes database by using a buffer around each pipe to take into account the small shifting that may arise between nodes and pipes. This ensures that the connections between pipes and nodes are accurately identified. After detecting the intersections, the resulting nodes are connected and inserted into the graph. During this phase, a verification is performed to ensure that intersections occur at the extremities of each pipe. If no intersections are found, dummy nodes are introduced to ensure the connectivity of the graph. This phase is consistent with the domain constraints (see section 3.1) where the nodes must be located at the extremities of the pipes.

The overall output of this process is a graph-based representation that accurately depicts the relationships between the pipes and nodes in the wastewater network. This process is detailed in the following section.

4 Graph construction

The graph construction algorithm (Algorithm 1) initializes an empty graph G with sets of nodes V and edges E. It iterates over each pipe p_i in the pipes database PD. For each pipe p_i, the algorithm follows five general steps detailed below:

Algorithm 1 Graph Construction

```
1: procedure GraphConstruction(PD, ND)
2:     V, E ← {}, {}
3:     G ← (V, E)
4:     Distance ← 0
5:     for i = 1 to N_P do
6:         Distance_{p_i} ← BufferDistance(p_i)
7:         Buffer_{p_i} ← CreateBuffer(p_i, Distance_{p_i})
8:         Intersections_{p_i} ← ExtractIntersections(ND, Buffer_{p_i})
9:         Ext1 ← RandomExtrimity(p_i)
10:        SortedInter_{p_i} ← DistanceAndSort(Intersections_{p_i},Ext1)
11:        SortedInter_{p_i} ← AddDummyNodes(SortedInter_{p_i})
12:        G ← ConnectionToGraph(SortedInter_{p_i}, G)
13:    return G
```

4.1 Lines/nodes connection

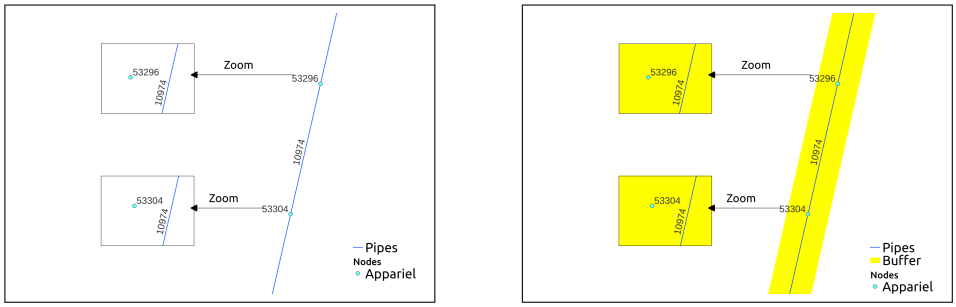
Step 1: Creation of a Buffer (line number 7 of Algorithm 1). In the GIS context, pipes are often represented using lines, which may not be entirely accurate. A pipe, being a substantial object, possesses attributes such as diameter. Hence, it is evident that there exists a distinction between pipes and simple line representations in GIS.

A buffer is created around each pipe, with its size reflecting the diameter of the pipe. As a result, detecting intersections between these buffers and nodes is synonymous with identifying intersections between pipes and nodes. From a mathematical perspective, a line technically has no diameter, meaning that a node is either precisely on the line or not. However, in practice, even minor numerical variations can arise, requiring the addition of a "margin" to ensure precision.

Inspired by GIS buffers, a buffer is defined as an area around a geographic feature containing all points within a specified distance from that feature. In this step, a buffer is created around each pipe p_i by applying a distance that accounts for the physical dimensions of the pipe and potential shifting of nearby nodes.

For example, in Figures 4a and 4b, the blue line represents the pipe, and a buffer is created around it using a distance. The resulting yellow buffer represents

the extended region around the pipe that encompasses its potential location within the given distance. It is clear from Figure 4a that there is no intersection between line number 10974 and the two nodes number 53296 and 53304 after zooming. However, in Figure 4b, we can see that there is an intersection between the yellow buffer and these nodes.



(a) (b)

Fig. 4: a) Intersection with GIS representation, b) Intersection using a buffer.

As the buffers represent the real dimensions of the pipes, their distances are exactly the diameter of the pipes. For example, the buffer distance shown in figure 5 is chosen based on the diameter of the pipe as presented in the algorithm 2. In the case where the diameter attribute in the GIS attribute table is empty or equal to zero, which is unrealistic, a default value is applied. This default value is set to the most frequently used diameter in the pipes database.

Algorithm 2 Calculate Buffer Distance

1: **procedure** BUFFERDISTANCE(p_i, frequentDiameter)
2: $DefaultDiameter \leftarrow$ frequentDiameter
3: **if** $Diameter_{p_i} \neq \text{Null} \wedge Diameter_{p_i} \neq 0$ **then**
4: $Distance_{p_i} \leftarrow \frac{Diameter_{p_i}}{2}$
5: **else**
6: $Distance_{p_i} \leftarrow \frac{DefaultDiameter}{2}$
7: return $Distance_{p_i}$

Step 2: Extraction of intersections (line number 8 of Algorithm 1). In the real-world, wastewater networks' pipes are connected to nodes. "Step 2" of the algorithm aims to detect intersections between the buffer created around the pipe p_i in "step 1" and the nodes databases ND in order to identify their

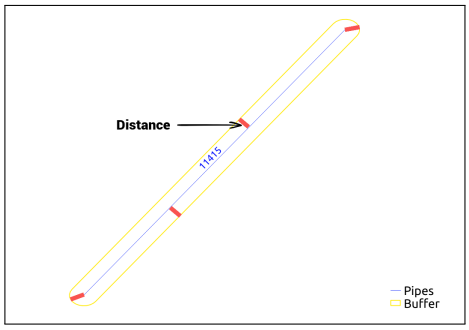


Fig. 5: Buffer creation

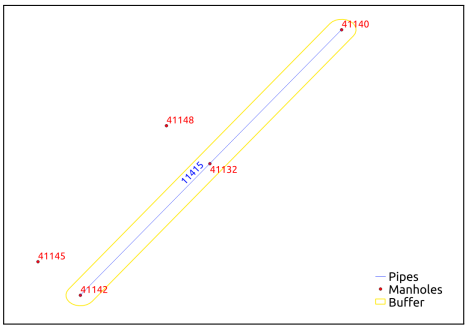


Fig. 6: Intersections detection

connections. These intersections between a pipe p_i and nodes database ND can be defined formally as follow:

$$\text{Intersections} = \{n_j \mid n_j \in \text{ND}, \exists s_l \in p_i, \text{shortestDistance}(n_j, s_l) \leq Distance_{p_i}\}$$

The set of intersections includes all nodes in proximity to the line, determined by the shortest distance, which is the Euclidean distance as explained later. The maximum allowable distance is defined as $Distance_{p_i}$.

The result is a set of nodes that must be connected to the corresponding pipe p_i. For example, in Figure 6, the intersection between the buffer of pipe 11415 and the ND is detected, resulting in a set of nodes denoted as Intersections = {41132, 41142, 41140}.

4.2 Lines extremities identification

Step 3: Computing distances and sorting intersections (line number 10 of Algorithm 1). In "step 3" of the algorithm, the Euclidean distance is calculated between one of the pipe's extremity $Extr1$ and each of the nodes intersecting the buffer (see algorithm 3). The Euclidean distance is the straight-line distance between two points in a two-dimensional space; it is suitable for cases where the distances between points are relatively small on surface Earth, which is the case in our situation with nodes that are near to each other in a wastewater network. Using Euclidean distance, the proximity of each intersection node to one of the extremities of the pipe, labeled as $Extr1$, can be determined. Sorting the intersection nodes based on their distances allows us to establish an order that reflects their positions relatively to the pipe's extremities.

Step 4: Add dummy nodes (line number 11 of Algorithm 1). This step is crucial for ensuring the connectivity of wastewater networks (see algorithm 4). To construct graphs based on the intersections, it is essential to have a node close to each extremity of the pipe p_i. However, this condition is not always satisfied. For instance, the set of intersections can be empty ($|SortedInter_{p_i}| = 0$ in

Algorithm 3 Calculate Euclidean Distance and Sort Intersections

procedure DistanceAndSort($Intersections_{p_i}$, $Extr1$)
  $distances \leftarrow \{\}$
  for *node* in $Intersections_{p_i}$ **do**
    $x, y \leftarrow$ coordinates of *node*
    $distance \leftarrow \sqrt{(x - Extr1.x)^2 + (y - Extr1.y)^2}$
    Append *distance* to *distances*
  $SortedInter_{p_i} \leftarrow$ Sort intersections based on *distances*
  return $SortedInter_{p_i}$

algorithm 4), or some nodes may be present but not located at the extremities of the pipe p_i.

In such scenarios, it becomes necessary to add dummy nodes to the extremities of the pipe. These dummy nodes serve as virtual representations of the missing nodes, facilitating the connection of the pipe to the overall wastewater network.

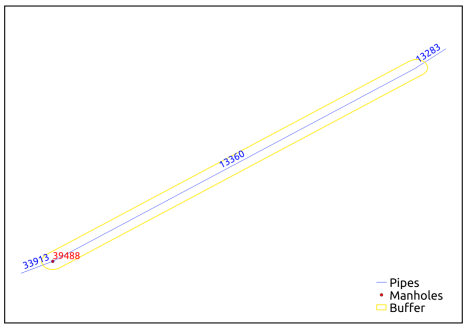


Fig. 7: Example of a pipe with one node intersecting the buffer area.

In Figure 7, it can be observed that there is only one node inside the buffer around pipe number 13360, which is near one extremity of the pipe. However, there is no node near the other extremity. To maintain connectivity with other pipes, a dummy node is necessary for the missing extremity.

4.3 Graph insertion (line number 8 of Algorithm 1).

The final step involves connecting the sorted intersections into the graph. After completing the step 4, the sorted intersections set is guaranteed to have at least two nodes that must be connected to the extremities of the pipe p_i. These nodes can either be from the original node set ND or dummy nodes that were added to ensure connectivity.

Algorithm 4 Adding Dummy Nodes

procedure AddDummyNodes($SortedInter_{p_i}$, p_i, $Distance_{p_i}$)
  $Ext1 \leftarrow$ the first extremity of p_i
  $Ext2 \leftarrow$ the second extremity of p_i
  if $|SortedInter_{p_i}| = 0$ **then**
    $SortedInter_{p_i} \leftarrow AddFirst(Type = Dummy, coord = (Ext1))$
    $SortedInter_{p_i} \leftarrow AddLast(Type = Dummy, coord = (Ext2))$
  else if $|SortedInter_{p_i}| = 1$ **then**
    $node \leftarrow$ node in $SortedInter_{p_i}$
    $distance1 \leftarrow EuclideanDistance(node, Ext1)$
    $distance2 \leftarrow EuclideanDistance(node, Ext2)$
    if $distance1 \geq Distance_{p_i}$ and $distance2 \geq Distance_{p_i}$ **then**
      $SortedInter_{p_i} \leftarrow AddFirst(Type = Dummy, coord = (Ext1))$
      $SortedInter_{p_i} \leftarrow AddLast(Type = Dummy, coord = (Ext2))$
    else if $distance1 \geq Distance$ **then**
      $SortedInter_{p_i} \leftarrow AddLast(Type = Dummy, coord = (Ext1y))$
    else if $distance2 \geq Distance$ **then**
      $SortedInter_{p_i} \leftarrow AddFirst(Type = Dummy, coord = (Ext2))$
  else if $|SortedInter_{p_i}| >= 2$ **then**
    $firstNode \leftarrow$ first node in $SortedInter_{p_i}$
    $lastNode \leftarrow$ last node in $SortedInter_{p_i}$
    if $EuclideanDistance(firstNode, Ext1) > Distance_{p_i}$ **then**
      $SortedInter_{p_i} \leftarrow AddFirst(Type = Dummy, coord = (Ext1))$
    if $EuclideanDistance(lastNode, Ext2) > Distance_{p_i}$ **then**
      $SortedInter_{p_i} \leftarrow AddLast(Type = Dummy, coord = (Ext2))$
  return $SortedInter_{p_i}$

The algorithm 5 starts by selecting the first node in the set of sorted intersections as the source node and adds it to the set of vertices in the graph G. Next, the algorithm iterates through the remaining nodes in this set. For each node, it creates an edge between the current source node and the destination node.

Algorithm 5 Connecting to the graph

procedure ConnectionToGraph($SortedInter$, G)
  $sourceNode \leftarrow$ first node in $SortedInter$
  $G.V \leftarrow G.V \cup \{sourceNode\}$
  for $node$ in $SortedInter$ excluding the first node **do**
    $destinationNode \leftarrow node$
    $E \leftarrow E \cup \{(sourceNode, destinationNode)\}$
    $sourceNode \leftarrow destinationNode$
  return G

4.4 Graph Construction illustration

In the illustrated example (Figure 8), the buffer zone created around pipe number 11415 is taken into consideration. This buffer intersects three nodes, namely Intersections = {41142, 41132, 41140}. These nodes represent the necessary connections to be established with pipe 11415. Before establishing these connections, a sorting process is performed and intersections with pipe extremities are verified. In this case, the intersections adequately cover the extremities, thus avoiding the need to introduce dummy nodes. Consequently, the graph can be constructed appropriately by connecting the sorted nodes to pipe 11415. It can be observed that a single pipe can result in the creation of two edges in the graph, as evidenced here with edges (41142, 41132) and (41132, 41140), owing to the connection of the middle node numbered 41132 to the pipe.

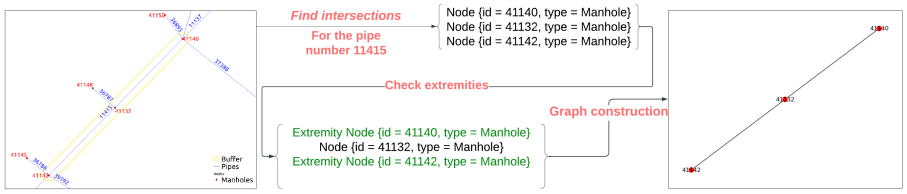


Fig. 8: Application of the algorithm: example with three nodes intersecting the buffer area including the pipe's extremities.

In the second example (Figure 9), involving pipe number 13360, a different situation arises. Only one intersection is found between this pipe and the nodes database, which happens to be located at one extremity. Unfortunately, no intersection exists at the other extremity. To ensure the proper connectivity of pipe 13360 within the graph, a dummy node with the label 107965 is introduced. This dummy node acts as a placeholder, allowing the inclusion of the missing extremity in the graph representation. Therefore, the resulting graph reflects the connection between node 39488 and the dummy node 107965, representing pipe 13360.

From the GIS-based representation of the wastewater network presented in Figure 1, a graph-based representation of these data is shown in Figure 10. Different colors of nodes symbolize specific types of components, with red indicating manholes. The edges represent the pipes, and their interconnections depict the real-world relationships found in wastewater networks. Furthermore, the graph-based representation retains all information about components from the GIS representation, encompassing both attribute and spatial details. This preservation enables us to generate GIS representations from the graph whenever required.

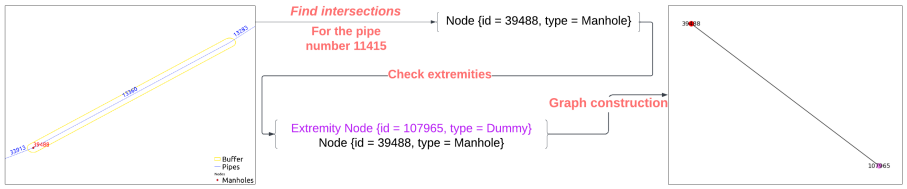


Fig. 9: Application of the algorithm: an example with three nodes intersecting the buffer area including the pipe's extremities.

5 Experiments

5.1 Datasets

The experiments involve five distinct real datasets of wastewater networks. These datasets cover the Montpellier Mediterranean Metropolis zone and originate from various sources and dates. The sources of these datasets are as follows:

- Veolia Company [20]:
  - Dataset 1: Year 2014.
  - Dataset 3: Year 2020.
- Montpellier Mediterranean Metropolis [14]:
  - Dataset 2: Year 2017.
  - Dataset 4: Year 2020.
  - Dataset 5: Year 2023.

Each dataset comprises various components of wastewater networks, with each component stored in a distinct database. Table 1 gives an overview of the different components present in each dataset, and their relative numbers. An hyphen (-) indicates that the component does not exist in the database.

datasets	1	2	3	4	5
Component					
DB1: Pipe	43789	47044	35531	37312	51295
DB2: Manholes	42007	45410	34371	37477	49987
DB3: Pumps	217	245	-	-	285
DB4: Structures	760	801	164	148	431
DB5: Treatment Plant	13	14	-	-	13
DB6: Accessories	-	-	108	198	-
DB7: Fittings	451	427	-	-	-
DB8: Devices	-	-	-	-	3808
DB9: Spillway	-	-	-	-	75

Table 1: Number of component in each database for the 5 datasets

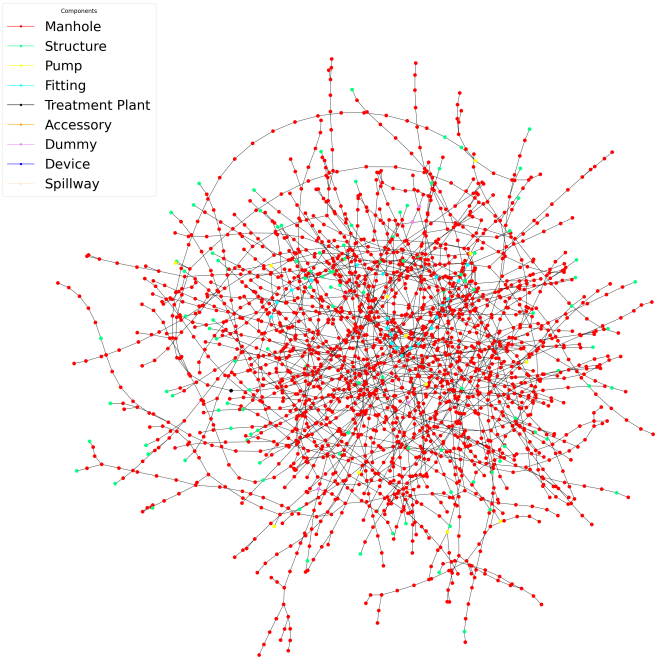


Fig. 10: Wastewater Graph-based representation derived from GIS-based representation of Figure 1

5.2 Pre-processing

The identification of duplicated objects can rely solely on spatial information, such as coordinates in our case, or both coordinates and attributes. It is evident that if multiple components share the same position and possess identical attributes, they are considered duplicates. In such instances, the need arises to retain only one representation.

Table 2a illustrates the number of duplicated objects for each component across all databases. A hyphen (-) indicates that the component does not exist in the database, while '0' signifies that the component's database exists but contains no duplicated objects. In other cases, it is the number of duplicated objects. Following the removal of duplicates based on all attributes, Table 2b presents the number of duplicated objects based solely on position. While these objects do not share the same values for all attributes, it remains impossible to have two manholes or pipes with the same geographical position. Notably, the majority of duplicated objects in our databases belong to the manholes and pipes components. Consequently, these repetitions must also be removed from the databases.

Datasets	1	2	3	4	5
Component					
DB1: Pipe	0	0	1	0	37
DB2: Manholes	0	0	65	0	29
DB3: Pumps	0	0	-	-	3
DB4: Structures	0	0	0	0	90
DB5: Treatment Plant	0	0	-	-	0
DB6: Accessories	-	-	0	0	-
DB7: Fittings	0	0	-	-	-
DB8: Devices	-	-	-	-	232
DB9: Spillway	-	-	-	-	0

(a) All attributes and position are identical

Datasets	1	2	3	4	5
Component					
DB1: Pipe	15	15	3	18	5
DB2: Manholes	0	1	65	206	32
DB3: Pumps	0	0	-	-	1
DB4: Structures	0	0	0	30	0
DB5: Treatment Plant	0	0	-	-	0
DB6: Accessories	-	-	0	0	-
DB7: Fittings	0	0	-	-	-
DB8: Devices	-	-	-	-	0
DB9: Spillway	-	-	-	-	0

(b) Position is identical but attributes differ.

Table 2: Duplicated Objects in Different Databases

5.3 Graph Construction

Table 3 presents the results of experiments conducted on the five datasets of wastewater networks. For each dataset, both the GIS representation and the graph-based representation derived from the GIS representation are available.

		datasets									
		1		2		3		4		5	
	representation	GIS	Graph	GIS	Graph	GIS	Graph	GIS	Graph	GIS	Graph
Components	DB1: Pipes	43789	44210	47044	47599	35531	35938	37312	38089	51295	55597
	DB2: Manholes	42007	41980	45410	45409	34371	34333	37477	37385	49987	49761
	DB3: Pumps	217	217	245	245	-	-	-	-	285	282
	DB4: Structures	760	760	801	801	164	164	148	105	431	424
	DB5: Treatment plants	13	13	14	14	-	-	-	-	13	11
	DB6: Accessories	-	-	427	427	108	74	198	168	-	-
	DB7: Fittings	451	451	-	-	-	-	-	-	-	-
	DB8: Devices	-	-	-	-	-	-	-	-	3808	3781
	DB9: Spillway	-	-	-	-	-	-	-	-	75	74
	Dummy nodes	-	405	-	189	-	1226	-	38	-	694

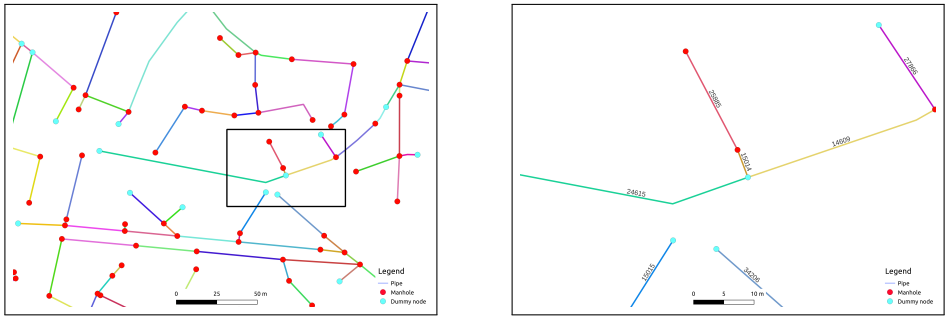
Table 3: From GIS to graph: Comparison of components number

For all datasets, most of the components from the GIS representation are successfully preserved in the graph. For instance, in dataset 1, the number of pipes in the GIS data is 43789, while the corresponding graph contains 44210 edges. It is expected that the number of edges in the graph will be equal to or greater than the number of pipes, as the algorithm maintains all pipes. In addition, in cases where the number of edges is greater than the number of pipes, the reason is that some pipes have been divided into several edges on the basis of their con-

nectivity with the nodes. The example shown in Figure 8 illustrates a scenario in which a single pipe is represented as two edges in the graph. Concerning the number of manholes, the GIS representation includes a total of 42007 manholes while the corresponding graph contains 41980. This represents a loss of 27 manholes, or around 0.064% of the total number of manholes. However, there are very few components missing in the graph representation. This happens when the components are located at a considerable distance from the nearest pipes in the GIS representation, making it challenging to establish direct connections between them. The algorithm ensures that each pipe in the GIS representation is represented as one or several edges in the graph-based representation.

5.4 Graph Connectivity

Figure 11a illustrates the integration of dummy nodes into the graph, highlighted in light blue. Each pipe is plotted in different color. These dummy nodes did not exist in the databases before the graph was constructed. Their introduction not only improves the connectivity of the network, it also plays an essential role in the completeness of the data. As the example shows, dummy nodes effectively simulate real-world objects that might otherwise be absent from the GIS database. in Figure 11a. Each pipe, shown in a different color, corresponds to a different entity in the pipes database. A zoom on the area inside the black rectangle is given in Figure 11b. It contains several pipes, but we focus on the small one numbered 15014, we can see that pipe 15014 has only one real node at one extremity, and no node at the other extremity. To add this pipe to the graph, two nodes are necessary at the extremities even if they are relatively close to each other. Therefore, adding a dummy node at the other extremity is necessary. This phase represents the initial step in improving the completeness of sewer network data.



(a) Added dummy nodes (b) Zoomed-in part

Fig. 11: Example of dummy nodes added in dataset 5

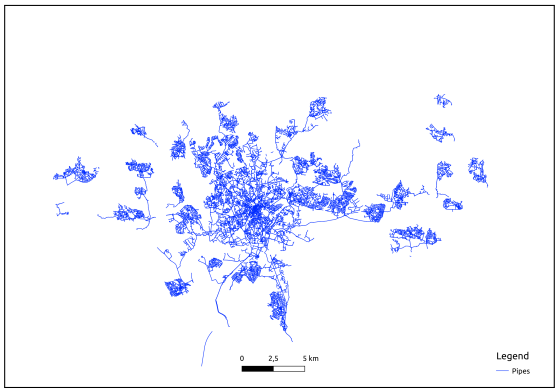


Fig. 12: General view of wastewater sub-networks in GIS representation of dataset 1

The number of connected sub-networks in the resulting graphs is an important criterion for evaluating the accuracy of the graph-based representation. In real-world wastewater networks (see figure 12), it is common to have many sub-networks due to various factors such as the physical layout of the network. However, in same cases the disconnected sub-networks in the GIS representation may be the result of insertion or shifting errors during data collection or digitization processes. These errors can lead to gaps in the network representation, making it challenging to accurately assess the connectivity of the wastewater network.

Dataset	1		2		3		4		5	
Representation	GIS	Graph	GIS	Graph	GIS	Graph	GIS	Graph	GIS	Graph
Number of Sub-networks	166	118	476	130	200	113	177	53	134	75

Table 4: Comparison of Sub-networks in GIS and Graph Representation

In the graph-based representation, one of the main goals is to identify and correct errors present in the GIS databases, which may lead to disconnected sub-networks in the resulting graph. The buffer distance used is designed to capture and address these errors of small shifting or inserting errors.

As a result of this error correction process, the number of sub-networks in the resulting graph is less than the number of sub-networks in the GIS representation (see Table 4). The reduction rate of the number of sub-networks depends on their distances from each other. For example, if sub-networks in a database are widely spaced, the reduction rate will be low. Conversely, if all sub-networks are close together in the database, the reduction rate will be high. In an ideal scenario, if each sub-network contains a node that is very near (in relation to the defined buffer distance) to a node in another sub-network, then all sub-networks will be

connected, forming a fully interconnected network. In our example, database 1 has fewer closely situated sub-networks compared to database 2. This discrepancy explains the differing rates of sub-network reduction. This reduction in the number of components indicates that the graph-based representation is successfully identifying and addressing the errors in the GIS representation, resulting in a more accurate and coherent representation of the wastewater network.

6 Conclusion

In this study, a graph-based approach for representing wastewater networks, derived from GIS databases, has been proposed. Our graph-based representation accurately captures the interconnected nature of the network. By conducting experiments on real-world datasets, the efficiency and effectiveness of the method have been demonstrated. It showcases the ability to preserve the maximum of components from the GIS representation while resolving missing connections between them.

As perspective, we aim to improve the graph-based representations by integrating water circulation within the pipes. This entails constructing oriented graph-based representations that reflect the flow direction and dynamics of wastewater through the network. Considering the direction of pipes allows us to further analyze flow behavior and optimize the network as a consequence. Furthermore, since all the datasets we possess represent the same geographical zone, our objective is to explore the fusion of these oriented graphs. The fusion process will enable us to synthesize a comprehensive representations that encompass the entire wastewater network for the given geographical area.

Acknowledgements This research has received support from the European Union's Horizon research and innovation program under the MSCA-SE (Marie Skłodowska-Curie Actions Staff Exchange) grant agreement 101086252; Call: HORIZON-MSCA-2021-SE-01; Project title: STARWARS (STormwAteR and WastewAteR networkS heterogeneous data AI-driven management).
This research has also received support from the French national project ANR (Agence Nationale de la Recherche) CROQUIS (Collecte, représentation, complétion, fusion et interrogation de données hétérogènes et incertaines de réseaux d'eaux urbains).
We would like to express our gratitude to "Montpellier Méditerranée Métropole" for having provided us with data essential to this research.

References

1. 3M. Guide technique eu 3m, 2024. Accessed on 2024-04-27.
2. Tanveer Ahmed, Khaista Rehman, Muhammad Shafique, and Wajid Ali. Gis-based earthquake potential analysis in northwest himalayan, pakistan. *Environmental Earth Sciences*, 82(4):113, 2023.

3. Peter A Burrough, Rachael A McDonnell, and Christopher D Lloyd. *Principles of geographical information systems*. Oxford University Press, USA, 2015.
4. Jonathan Campbell and Michael Shin. *Geographic information system basics*. 2012 Book Archive, 2012.
5. Jack Dangermond. A classification of software components commonly used in geographic information systems. *Design and implementation of computer-based geographic information systems*, pages 70–91, 1983.
6. Omar Et-Targuy, Ahlame Begdouri, Salem Benferhat, Carole Delenne, Thanh-Nghi Do, and Truong-Thanh Ma. A graph-based approach for representing wastewater networks from gis data: Ensuring connectivity and consistency. In *International Conference on Intelligent Systems and Data Science*, pages 243–257. Springer, 2023.
7. Andrew Fall, Marie-Josée Fortin, Micheline Manseau, and Dan O'Brien. Spatial graphs: principles and applications for habitat connectivity. *Ecosystems*, 10:448–461, 2007.
8. Darij Grinberg. An introduction to graph theory. *arXiv preprint arXiv:2308.04512*, 2023.
9. Robert J. Hijmans, Susan E. Cameron, Juan L. Parra, Peter G. Jones, and Andy Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965 – 1978, 2005. Cited by: 15799.
10. Mingfei Liu, Xinyu Li, Jie Li, Yahui Liu, Bin Zhou, and Jinsong Bao. A knowledge graph-based data representation approach for iiot-enabled cognitive manufacturing. *Advanced Engineering Informatics*, 51:101515, 2022.
11. Kang Mao, Hua Zhang, Yuwei Pan, and Zhugen Yang. Biosensors for wastewater-based epidemiology for monitoring public health. *Water research*, 191:116787, 2021.
12. Thomas Maugey, Yung-Hsuan Chao, Akshay Gadde, Antonio Ortega, and Pascal Frossard. Luminance coding in graph-based representation of multiview images. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 130–134. IEEE, 2014.
13. Thomas Maugey, Antonio Ortega, and Pascal Frossard. Graph-based representation for multiview image geometry. *IEEE Transactions on Image Processing*, 24(5):1573–1586, 2015.
14. Montpellier Méditerranée Métropole. Open data - montpellier méditerranée métropole. `https://data.montpellier3m.fr/`, 2024. Accessed on 2024-03-24.
15. Amare GebreMedhin Nigusse, Ukeubay Geiday Adhaneom, Gebrerufael Hailu Kahsay, Abadi Mehari Abrha, Desta Nigusse Gebre, and Amanuel Gedey Weldearegay. Gis application for urban domestic wastewater treatment site selection in the northern ethiopia, tigray regional state: a case study in mekelle city. *Arabian Journal of Geosciences*, 13:1–13, 2020.
16. Ali Reza Noori and SK Singh. Assessment and modeling of sewer network development utilizing arc gis and sewergems in kabul city of afghanistan. *Journal of Engg. Research ICARI Special Issue pp*, 22:31, 2021.
17. Abolghasem Sadeghi-Niaraki, Masood Varshosaz, Kyehyun Kim, and Jason J Jung. Real world representation of a road network for route planning in gis. *Expert systems with applications*, 38(10):11999–12008, 2011.
18. Sheetal S Sonawane and Parag A Kulkarni. Graph based representation and analysis of text document: A survey of techniques. *International Journal of Computer Applications*, 96(19), 2014.
19. Michael R Templeton and David Butler. *Introduction to wastewater treatment*. Bookboon, 2011.

20. Veolia. Veolia company. `https://www.veolia.fr/`, 2024. Accessed on 2024-04-27.
21. Michael Zeiler. *Modeling our world: the ESRI guide to geodatabase design.* ESRI, Inc., 1999.
22. Hao Zhou, Xu Yang, Mingyu Fan, Hai Huang, Dongchun Ren, and Huaxia Xia. Static-dynamic global graph representation for pedestrian trajectory prediction. *Knowledge-Based Systems*, 277:110775, 2023.